Molecular Spectroscopy

Liang Wang

Advisor: Prof. Liang Wang
Institute of Automation, Chinese Academy of Sciences
2 December 2024

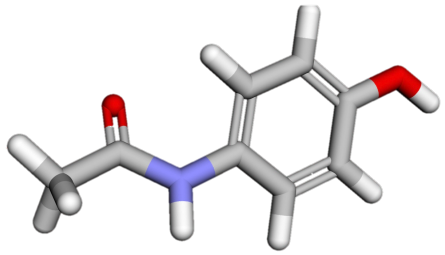
Outline

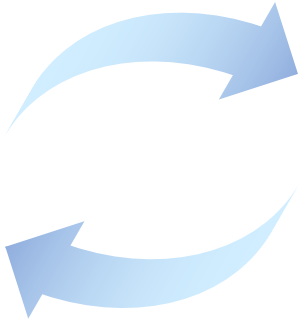
- Background
- Task 1: molecular spectrum prediction
  - A deep learning model for predicting selected organic molecular spectra, NCS 2023
  - Tandem mass spectrum prediction for small molecules using graph transformers, NMI 2024
- Task 2: molecular structure elucidation
  - End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction, arXiv 2024
- Recent datasets and benchmarks
  - Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry, NeurIPS 2024
  - MassSpecGym: A benchmark for the discovery and identification of molecules, NeurIPS 2024
  - Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation, NeurIPS 2024

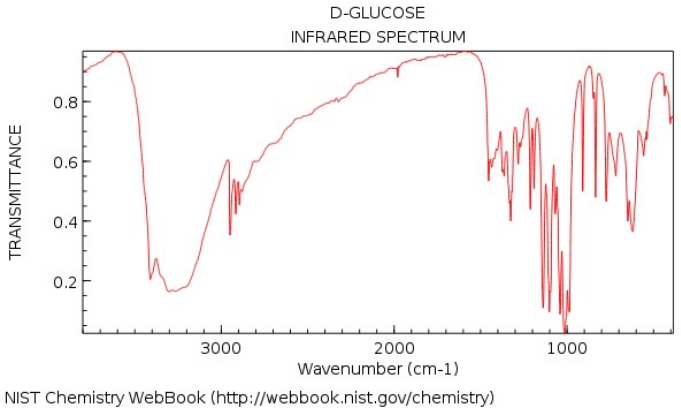
Introduction to Molecular Spectroscopy

- **Applications:** identify the structure of compounds

Task 1: Molecular Spectrum Prediction







Molecular Structure

Molecular Spectrum

Task 2: Molecular Structure Elucidation

Introduction to Molecular Spectroscopy

- **Definition:** The study of interactions between molecules and various forms of energy (e.g., electromagnetic radiation, magnetic fields) to gain insights into molecular structure, composition, and properties.

- **Importance:** Key tool in chemistry, physics, and biology for understanding molecular structure and dynamics.

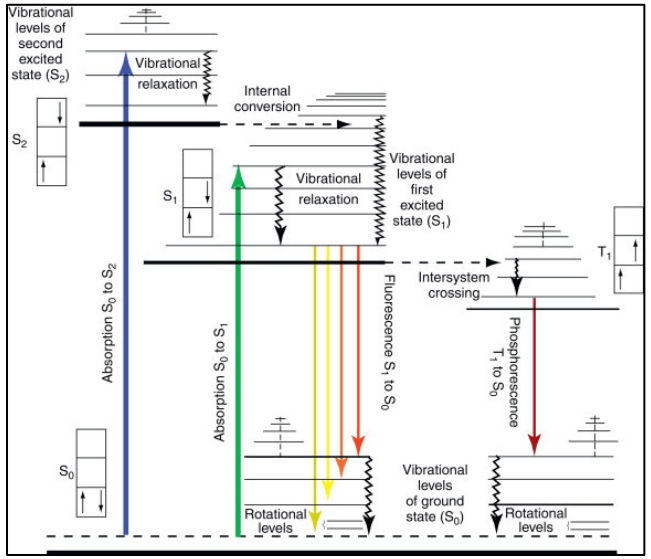
- **Applications:**
  - Identifying chemical compounds.

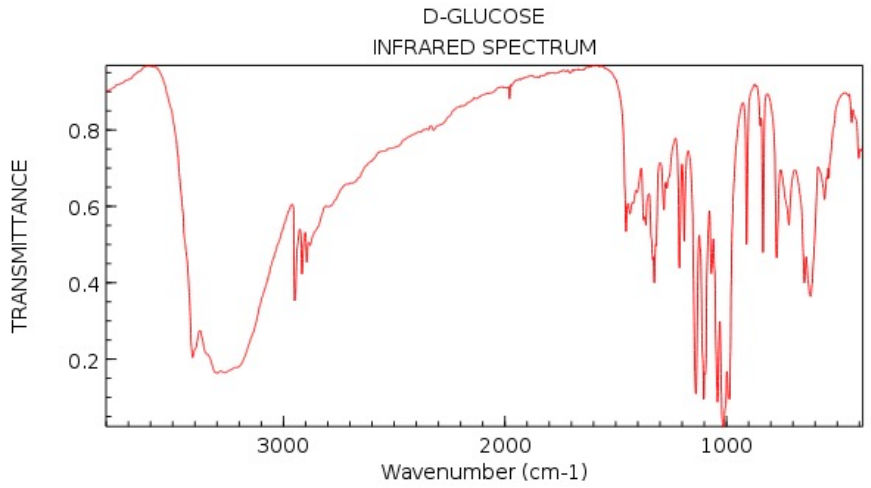
Introduction to Molecular Spectroscopy

- **Key Techniques**:
  - **1 Spectroscopic Techniques**:
    - *Infrared (IR)*: Analyzes molecular vibrations.
    - *Ultraviolet-Visible (UV-Vis)*: Studies electronic transitions.
  - **2 Magnetic Resonance Techniques**:
    - *Nuclear Magnetic Resonance (NMR)*: Provides structural information through nuclear spin interactions.
    - *Electron Paramagnetic Resonance (EPR)*: Examines unpaired electron spins.
  - **3 Diffraction Techniques**:
    - *X-ray Diffraction (XRD)*: Reveals 3D structures via X-ray scattering.
  - **4 Mass Spectrometry**:
    - Determines molecular weight and structure by analyzing ionized samples.

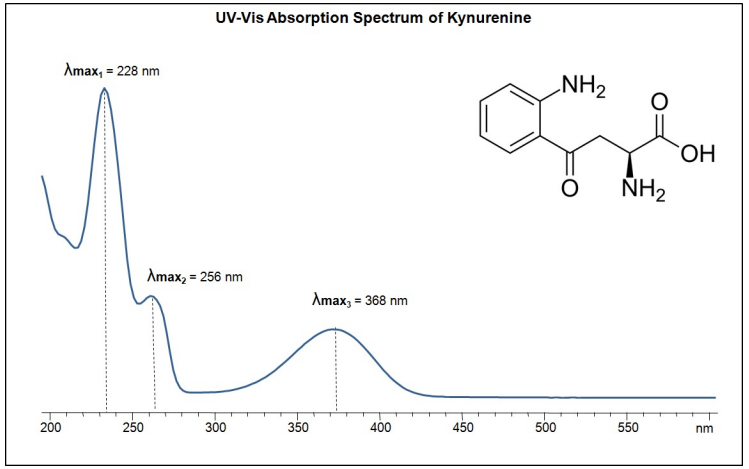
Introduction to Molecular Spectroscopy

- **Key Techniques**:
  - **1 Spectroscopic Techniques**:
    - *Infrared (IR)*: Analyzes molecular vibrations.
    - *Ultraviolet-Visible (UV-Vis)*: Studies electronic transitions.



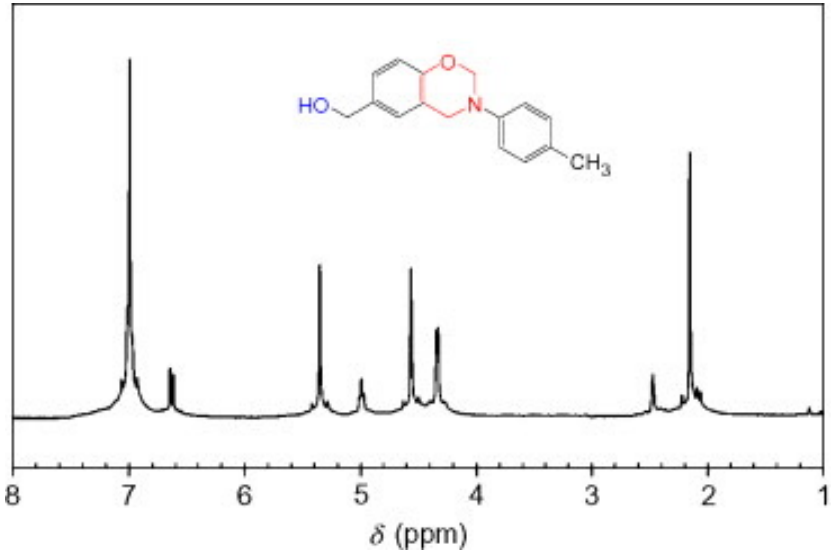


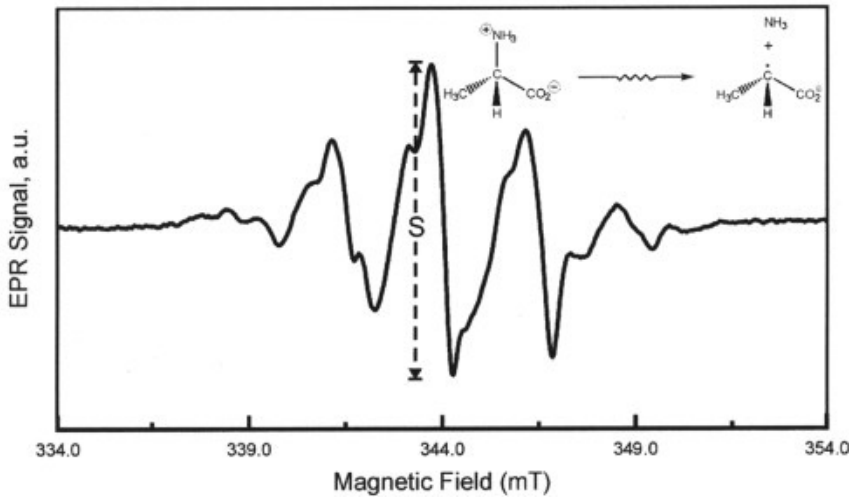
NIST Chemistry WebBook (http://webbook.nist.gov/chemistry)



Introduction to Molecular Spectroscopy

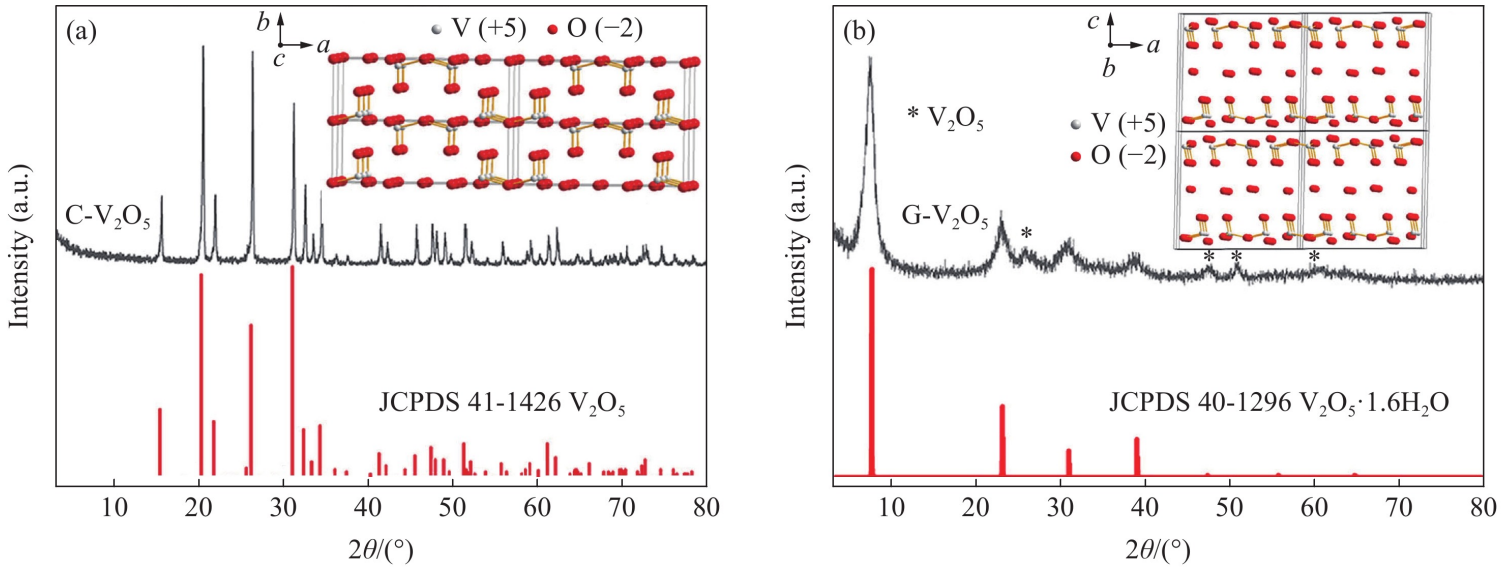
- **Key Techniques**:
  - **2 Magnetic Resonance Techniques**:
    - *Nuclear Magnetic Resonance (NMR)*: Provides structural information through nuclear spin interactions.
    - *Electron Paramagnetic Resonance (EPR)*: Examines unpaired electron spins.





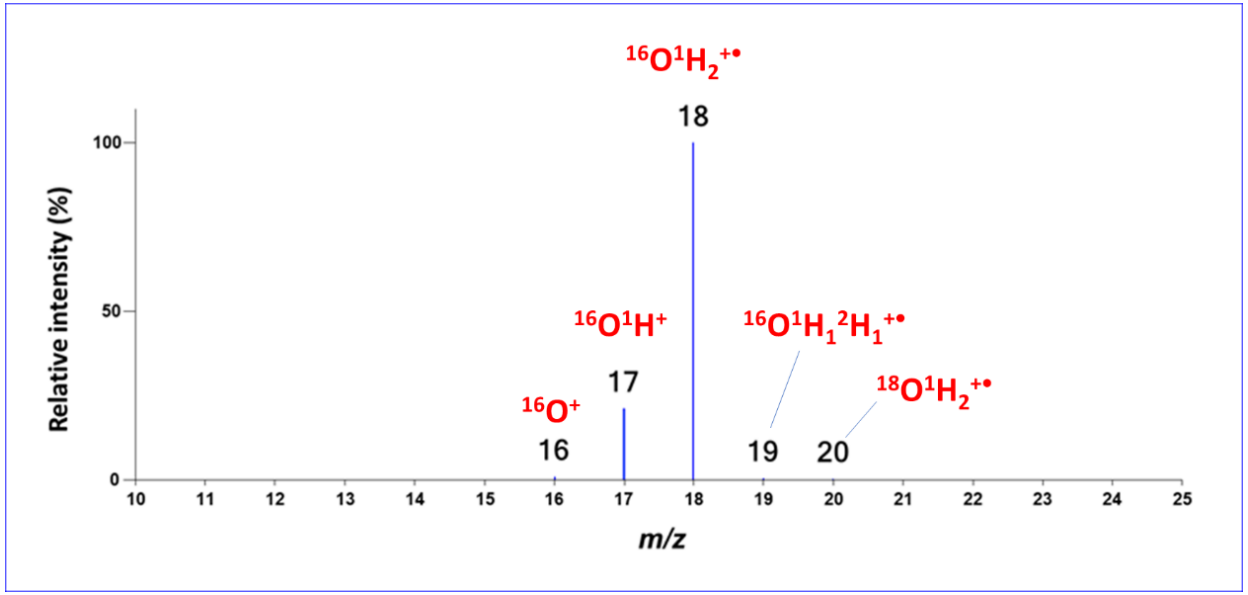
Introduction to Molecular Spectroscopy

- **Key Techniques**:
  - **3 Diffraction Techniques**:
    - *X-ray Diffraction (XRD)*: Reveals 3D structures via X-ray scattering.



Introduction to Molecular Spectroscopy

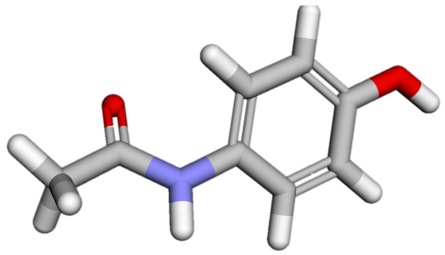
- **Key Techniques**:
  - **4 Mass Spectrometry**:
    - Determines molecular weight and structure by analyzing ionized samples.

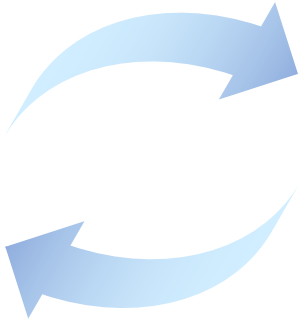


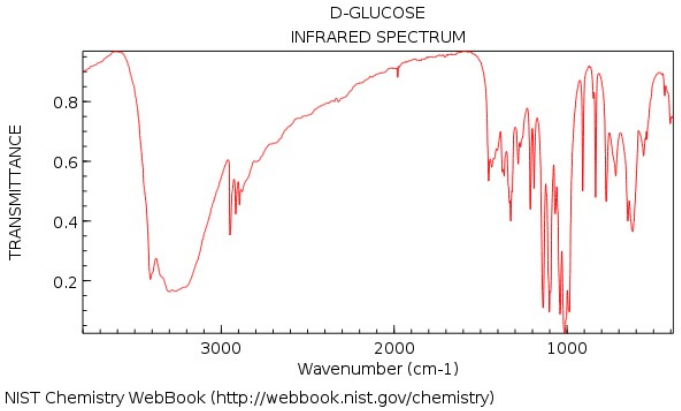
Introduction to Molecular Spectroscopy

- **Applications:** identify the structure of compounds

Task 1: Molecular Spectrum Prediction







Molecular Structure

Molecular Spectrum

Task 2: Molecular Structure Elucidation

Task 1: Molecular Spectrum Prediction

PaiNN, ICML 2021

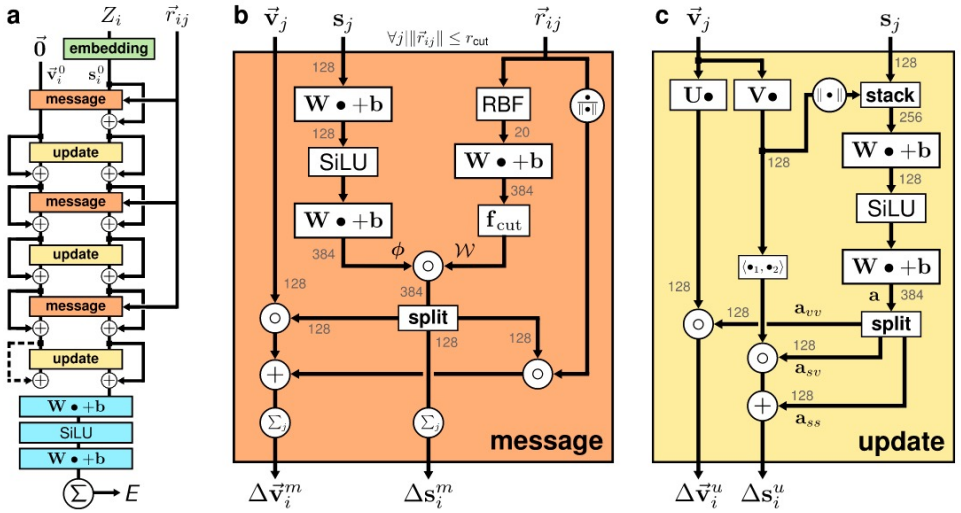
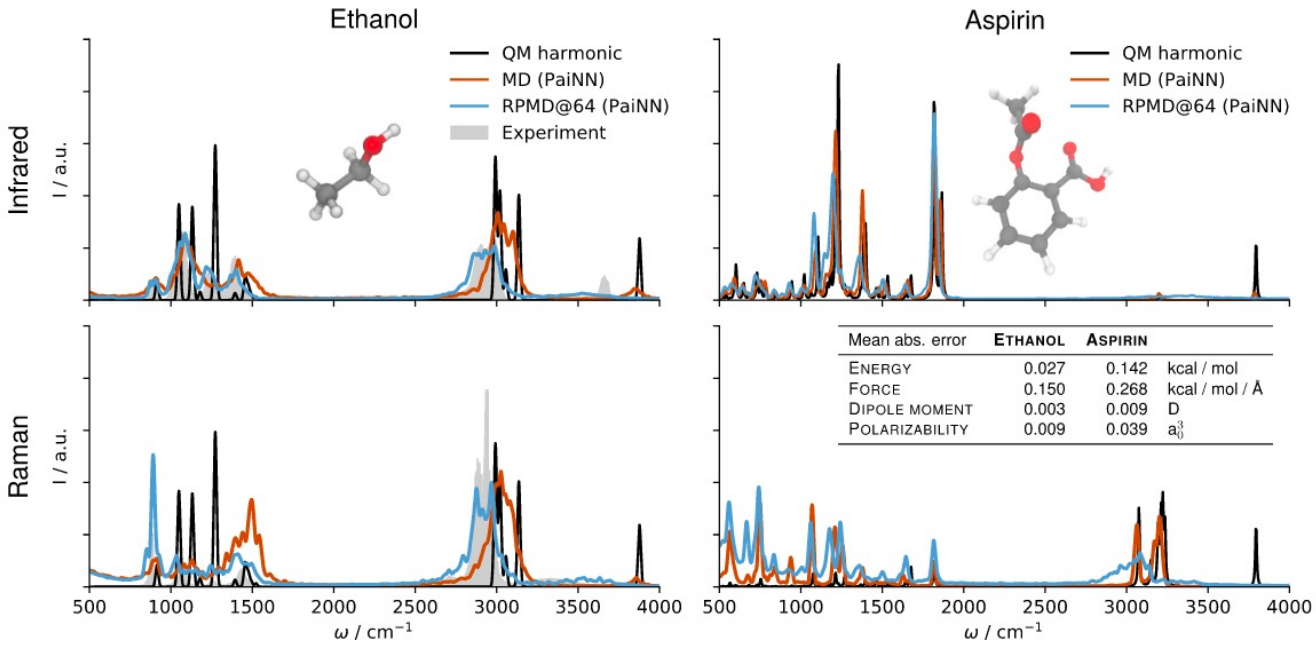


Figure 2. The architecture of PAINN with the full architecture (**a**) as well as the message (**b**) and update blocks (**c**) of the eq message passing. In all experiments, we use 128 features for $\mathbf{s}_i$ and $\vec{\mathbf{v}}_i$ throughout the architecture. Other layer sizes are annotate



dipole moment

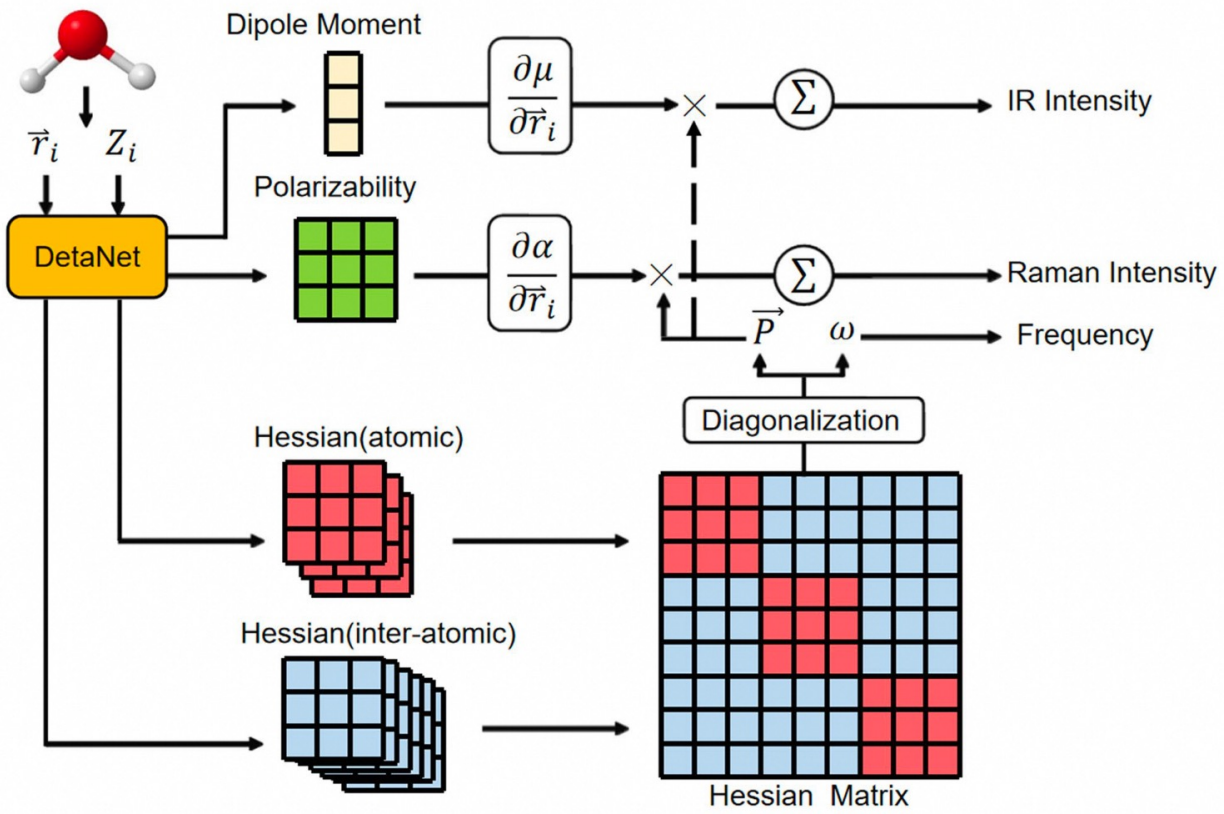
$$\vec{\mu} = \sum_{i=1}^{N} \vec{\mu}_{\text{atom}}(\vec{\mathbf{v}}_i) + q_{\text{atom}}(\mathbf{s}_i)\vec{r}_i.$$

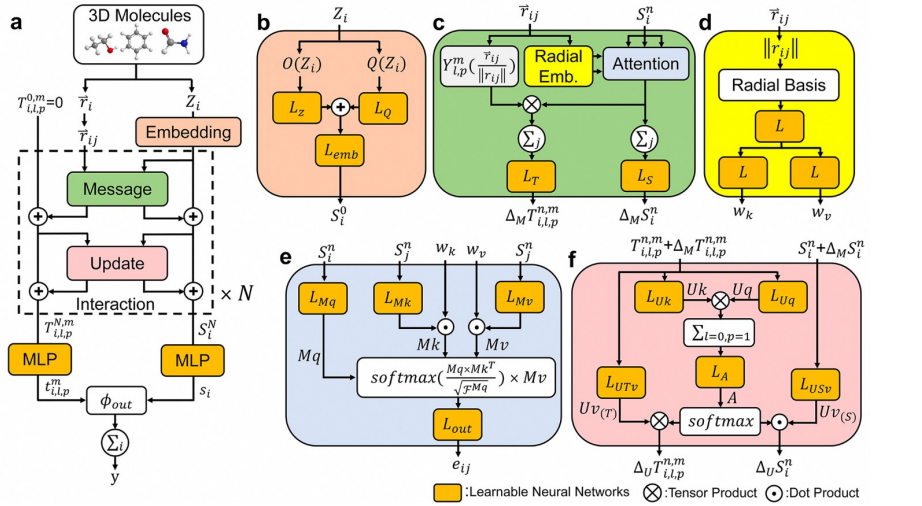
polarizability

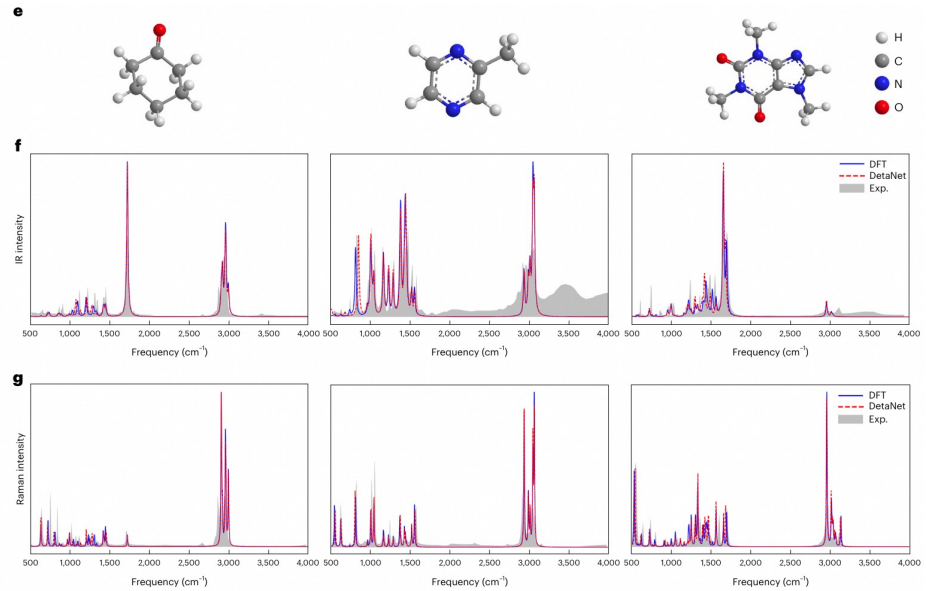
$$\boldsymbol{\alpha} = \sum_{i=1}^{N} \alpha_0(\mathbf{s}_i)\, I_3 + \vec{\nu}(\vec{\mathbf{v}}_i) \otimes \vec{r}_i + \vec{r}_i \otimes \vec{\nu}(\vec{\mathbf{v}}_i),$$

Equivariant message passing for the prediction of tensorial properties and molecular spectra, ICML 2021

DetaNet, NCS 2023

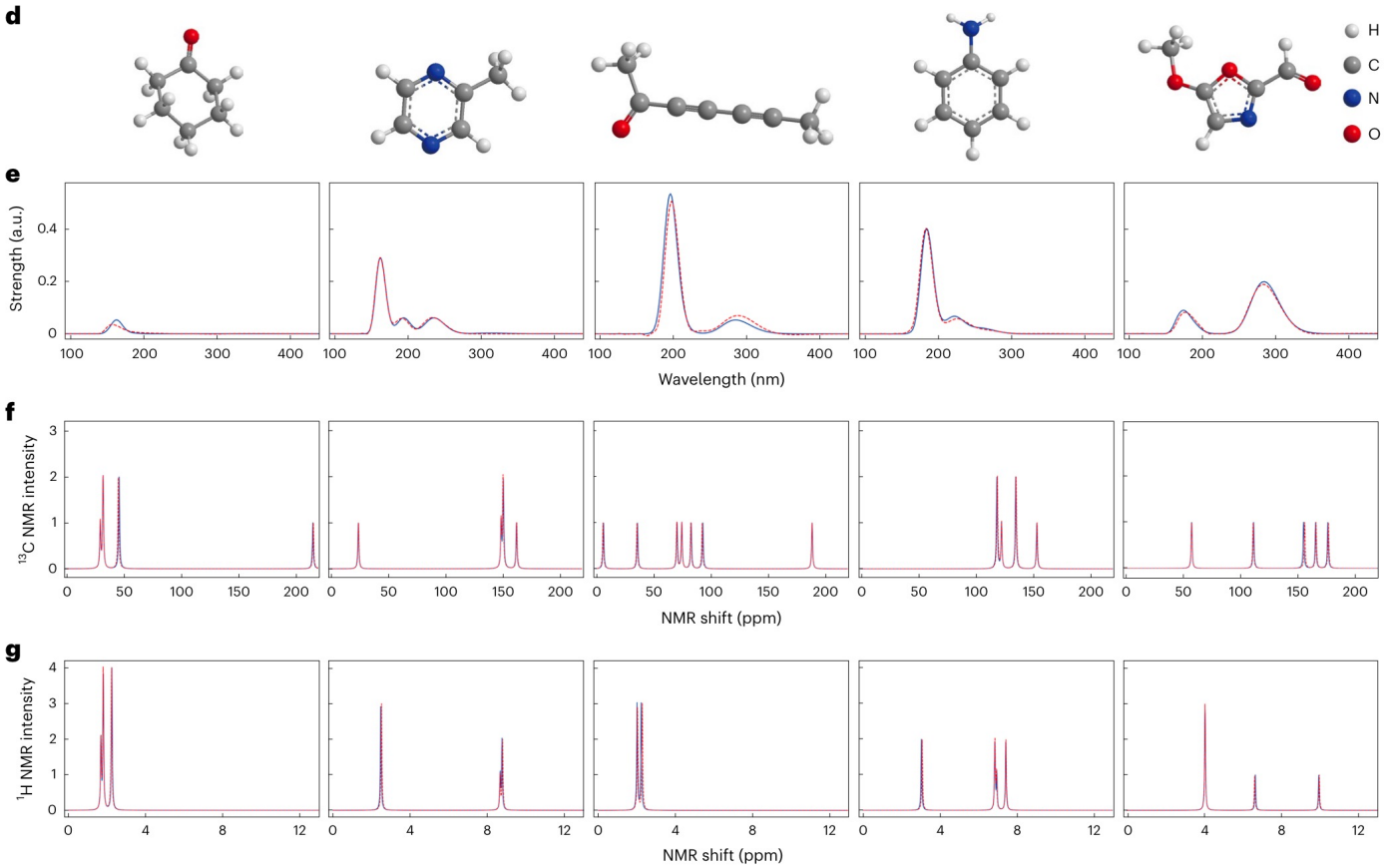


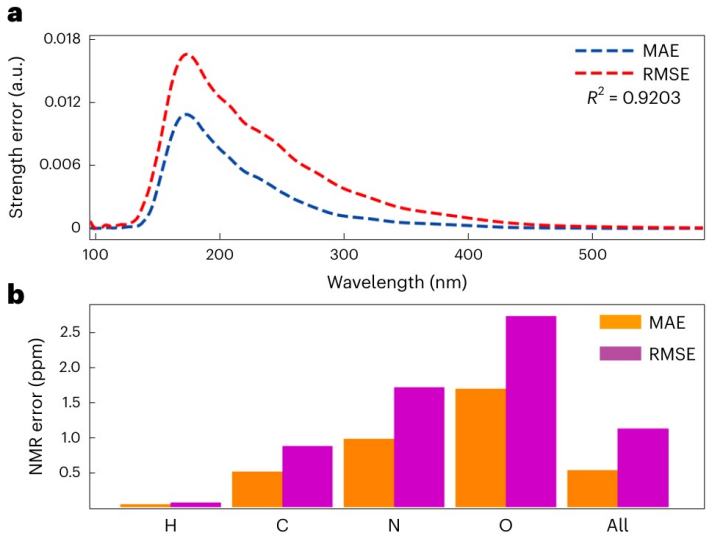


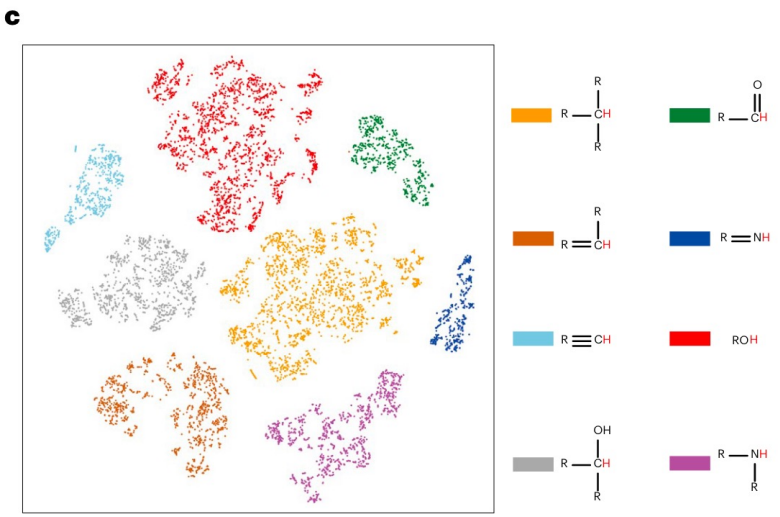


A deep learning model for predicting selected organic molecular spectra, NCS 2023-11-13

DetaNet, NCS 2023

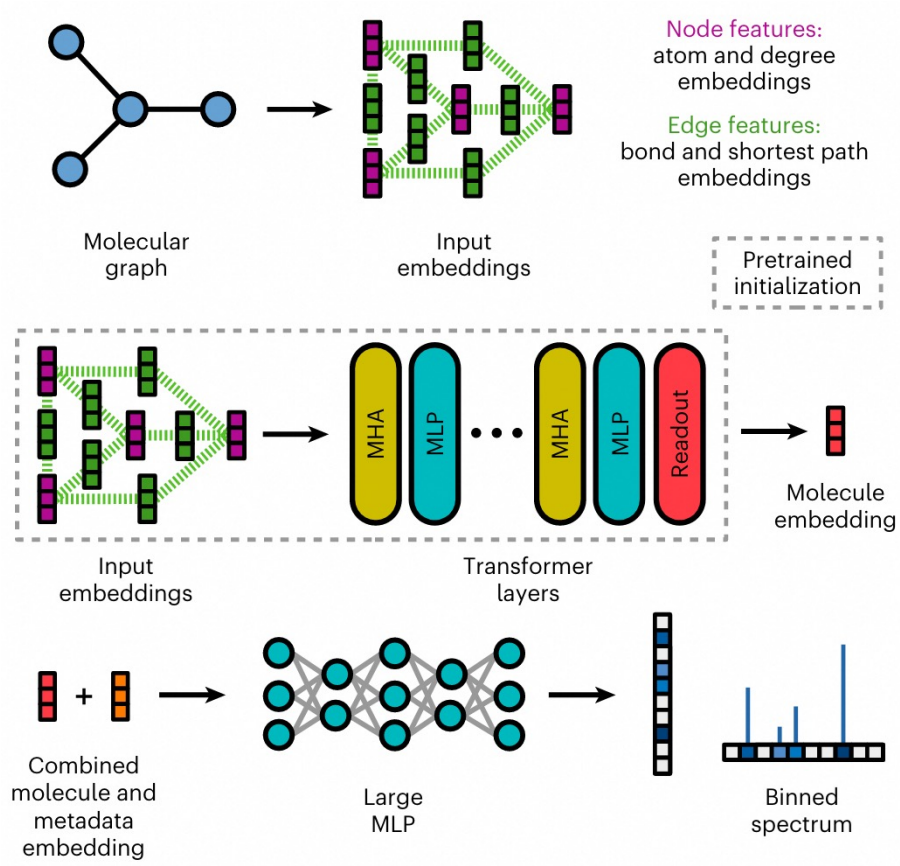






A deep learning model for predicting selected organic molecular spectra, NCS 2023-11-13

MassFormer, NMI 2024



MassFormer

- Graph transformer to model long-range relationships between atoms.
- The transformer modules are pre-trained through a chemical task (pre-trained Graphormer), then fine-tuned on spectral data.
- Divide the spectral vector by bin

$$\mathbf{s} = f_\psi\left(\mathbf{h}_{N+1}^{L} \parallel \mathbf{z}\right)$$

$$\mathrm{CD}(\mathbf{y}, \hat{\mathbf{y}}) = 1 - \frac{\mathbf{y}^T\hat{\mathbf{y}}}{\parallel \mathbf{y} \parallel_2 \parallel \hat{\mathbf{y}} \parallel_2} = 1 - \frac{\sum_{i=1}^{m} y_i \hat{y}_i}{\sqrt{\sum_{j=1}^{m} y_j^2 \sum_{k=1}^{m} \hat{y}_k^2}}$$

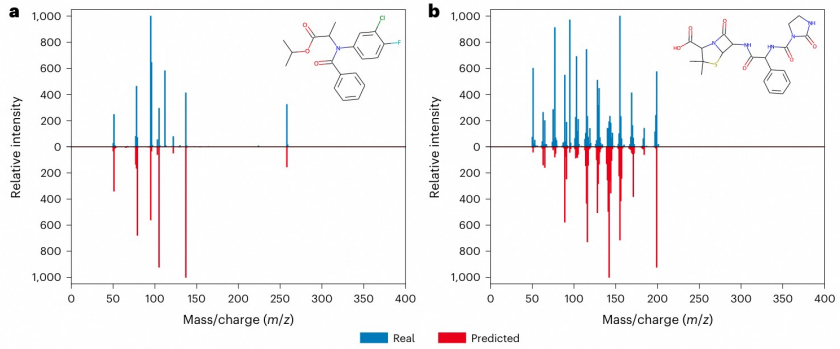
$$\hat{y}_{\mathrm{F}}(\mathbf{s})_i = \left(W_{\mathrm{F}}\mathbf{s} + \mathbf{b}_{\mathrm{F}}\right)_i$$

$$\hat{y}_{\mathrm{R}}(\mathbf{s})_{m_p+\tau-i} = \left(W_{\mathrm{R}}\mathbf{s} + \mathbf{b}_{\mathrm{R}}\right)_i$$

$$\hat{y}_{\mathrm{G}}(\mathbf{s})_i = \mathtt{sigmoid}\left[W_{\mathrm{G}}\mathbf{s} + \mathbf{b}_{\mathrm{G}}\right]_i$$

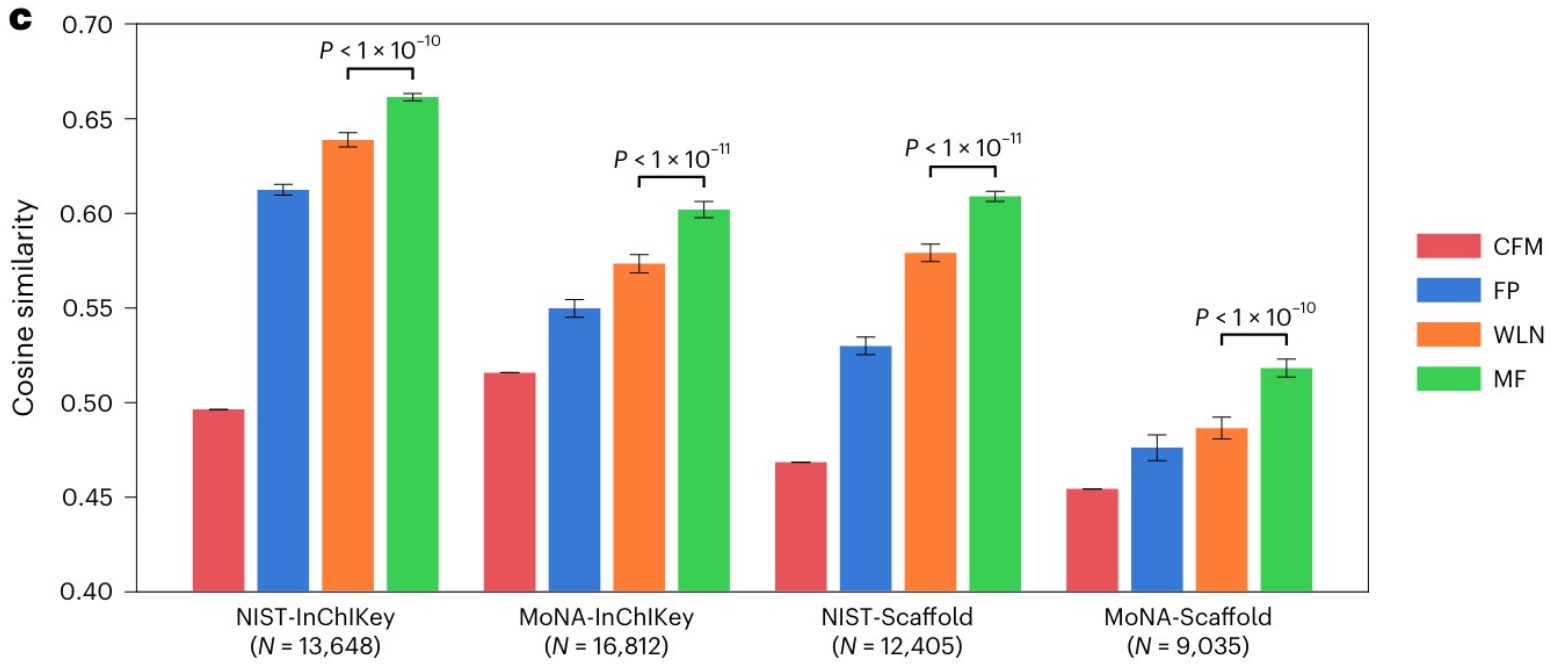
$$\hat{y}_{\mathrm{FR}}(\mathbf{s})_i = \hat{y}_{\mathrm{G}}(\mathbf{s})_i \hat{y}_{\mathrm{F}}(\mathbf{s})_i + (1 - \hat{y}_{\mathrm{G}}(\mathbf{s})_i)\hat{y}_{\mathrm{R}}(\mathbf{s})_i$$

$$\hat{y}(\mathbf{s})_i = \mathtt{relu}\left[\mathbb{I}\left[i \le m_p + \tau\right]\hat{y}_{\mathrm{FR}}(\mathbf{s})_i\right]$$



Tandem mass spectrum prediction for small molecules using graph transformers, NMI 2024-04-05

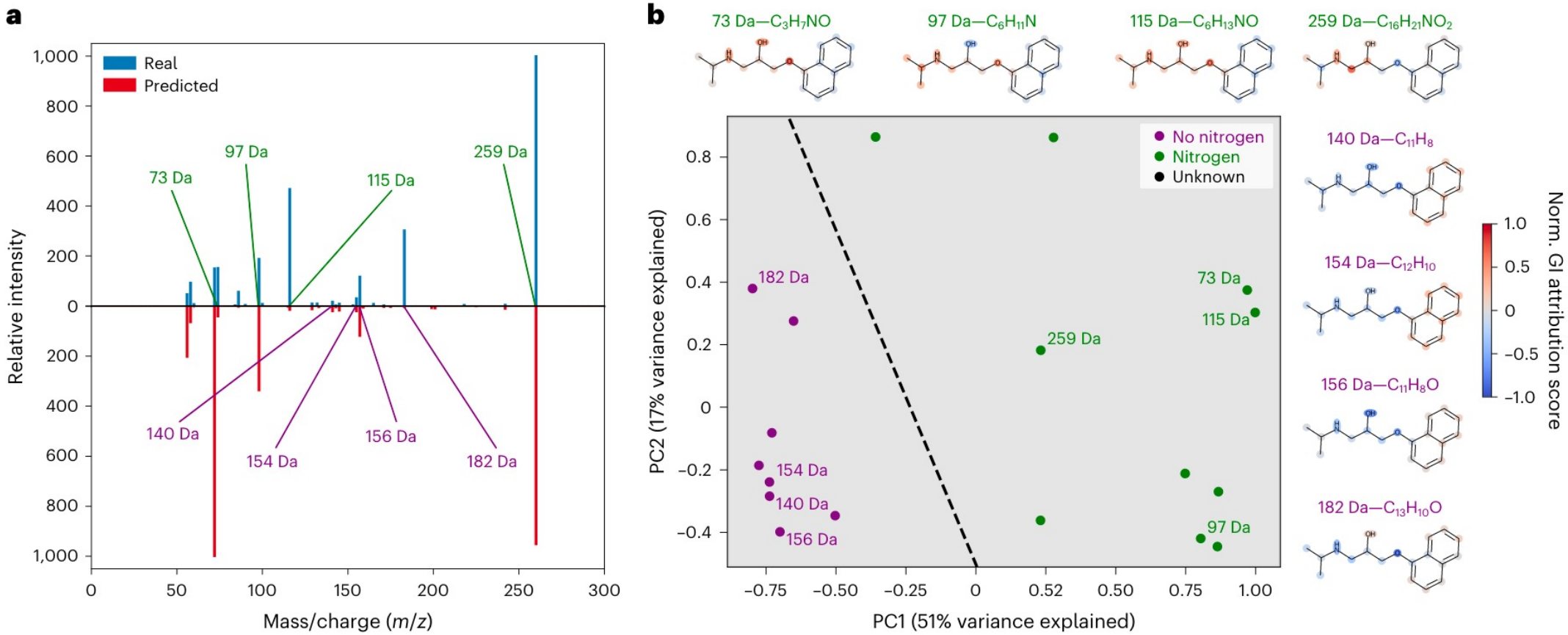
MassFormer, NMI 2024



Tandem mass spectrum prediction for small molecules using graph transformers, NMI 2024-04-05

MassFormer, NMI 2024

Explaining peak predictions with gradient-based attribution



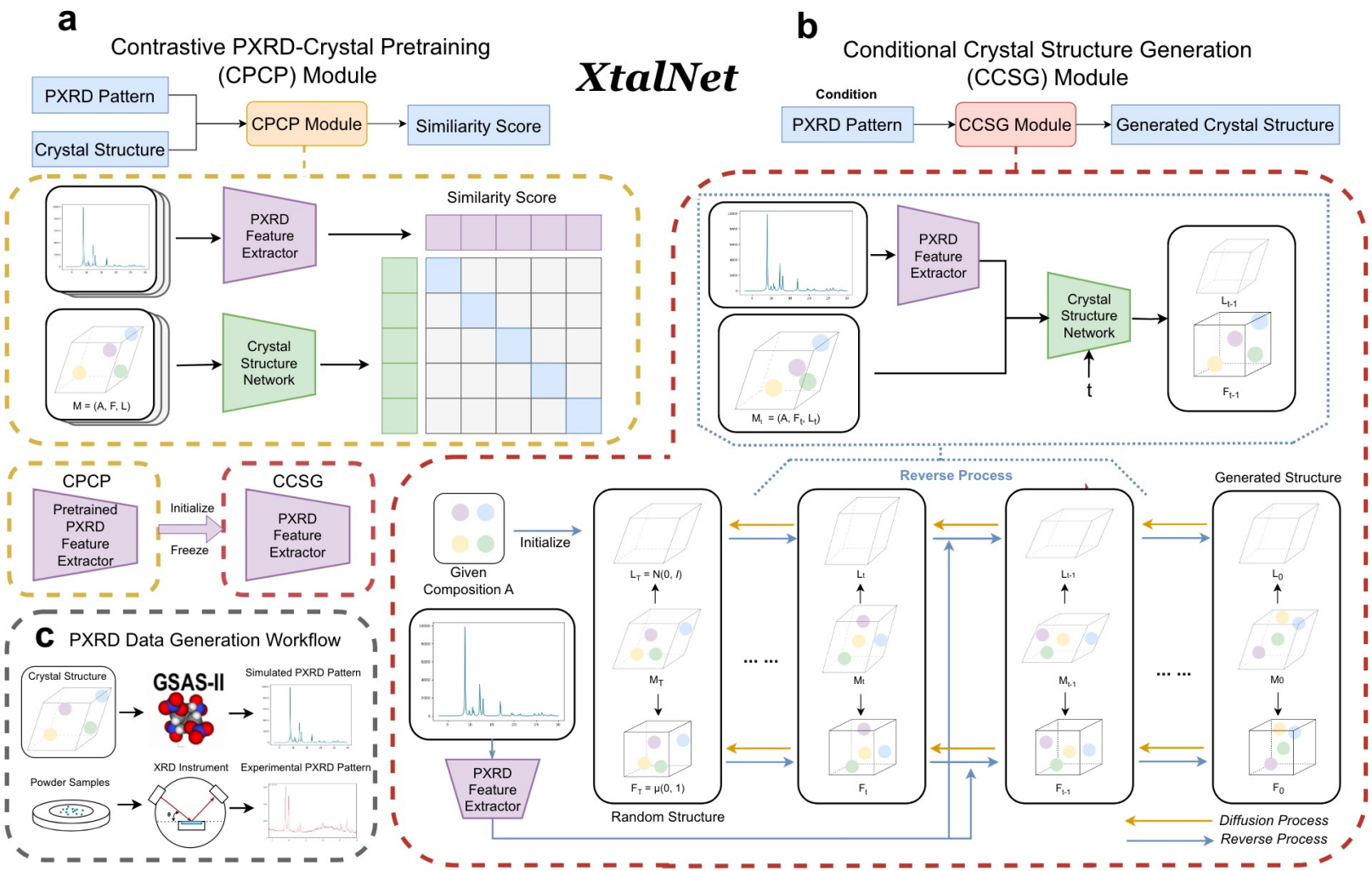
Tandem mass spectrum prediction for small molecules using graph transformers, NMI 2024-04-05

Task 2: Molecular Structure Elucidation

XtalNet, arXiv 2024

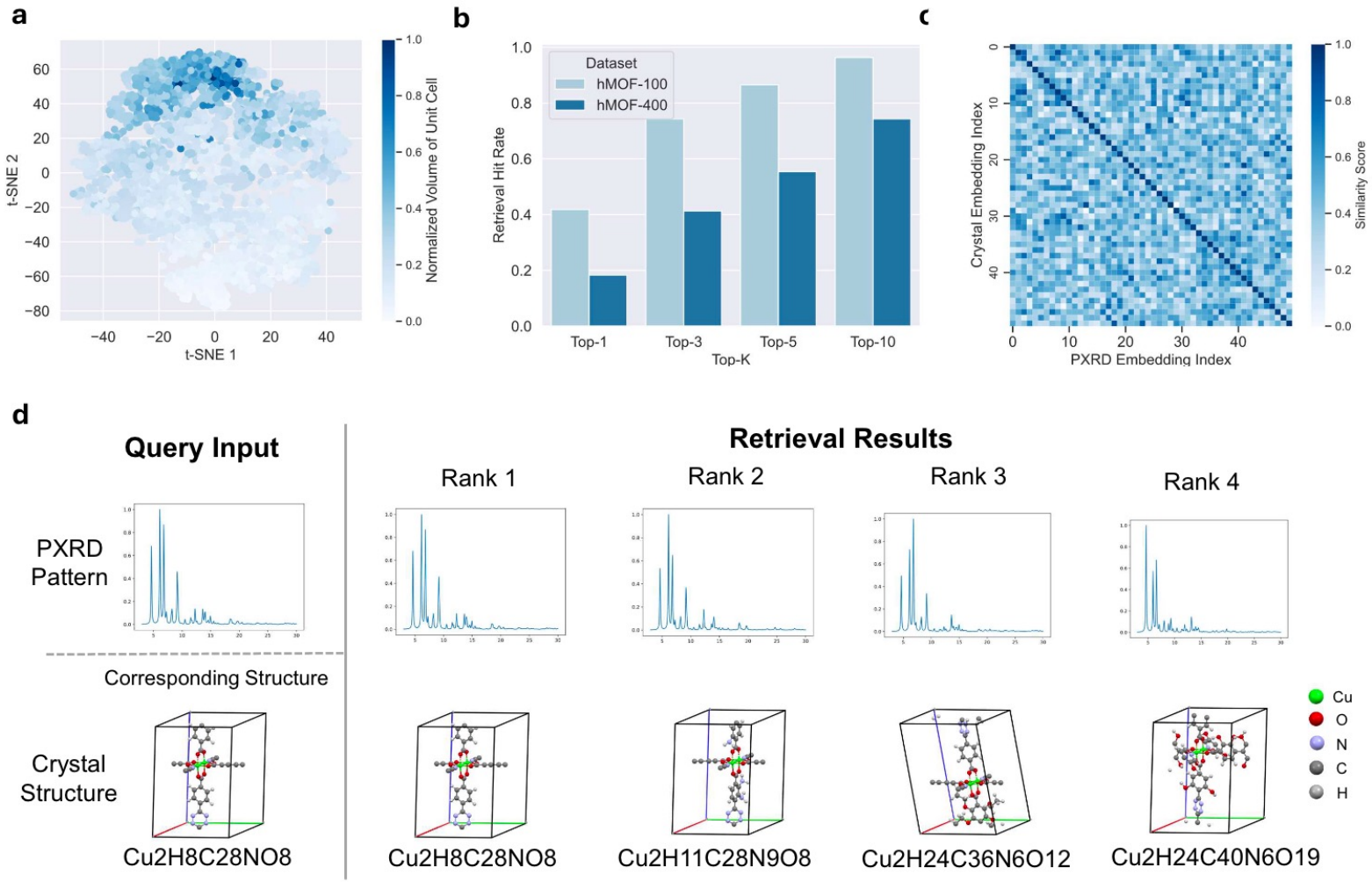
$$p(\boldsymbol{L}, \boldsymbol{F} \mid \boldsymbol{A}, \boldsymbol{C}_{XRD})$$

$$\boldsymbol{P} = f_{PXRD}(\boldsymbol{A}_{xrd}, \boldsymbol{I}_{xrd})$$



End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction, arXiv 2024-04

XtalNet, arXiv 2024



End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction, arXiv 2024

XtalNet, arXiv 2024

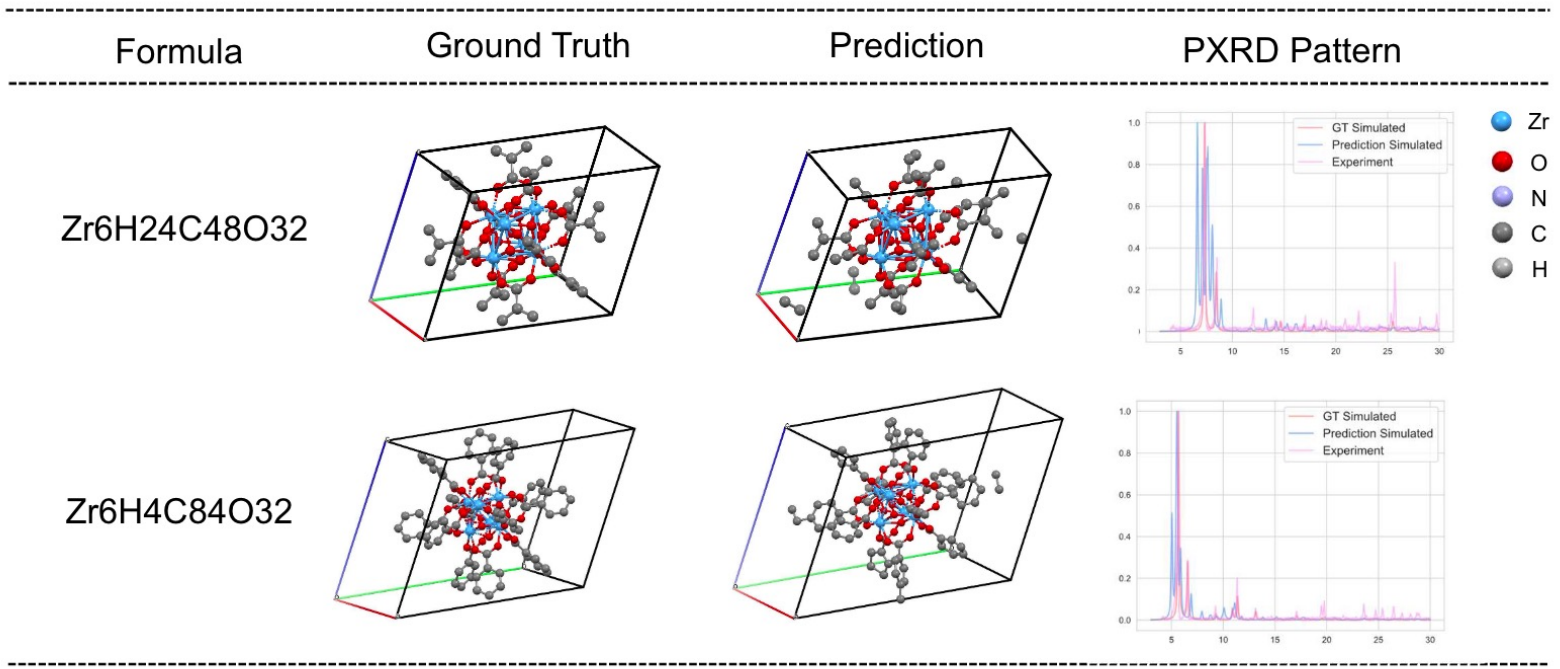


Fig. 5 XtalNet Predictions of Real Experimental PXRD Patterns. Two cases of XtalNet's predictions for real experimental PXRD data are drawn, showcasing both the ground truth (GT) crystal structures and the predicted crystal structures. The GT simulated (red), predicted simulated (blue), and experimental (purple) PXRD patterns are also presented for comparison.

End-to-End Crystal Structure Prediction from Powder X-Ray Diffraction, arXiv 2024

Datasets and Benchmarks

multi-modal spectra dataset, NeurIPS 2024

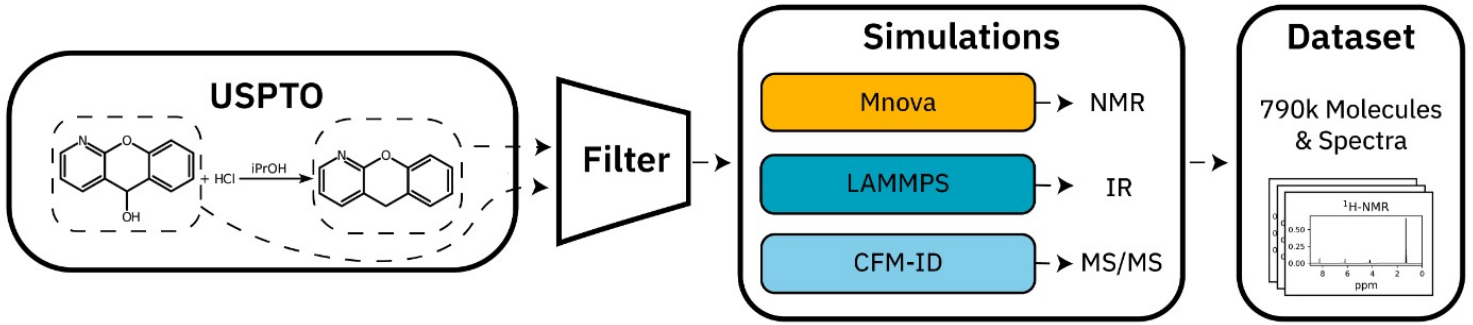
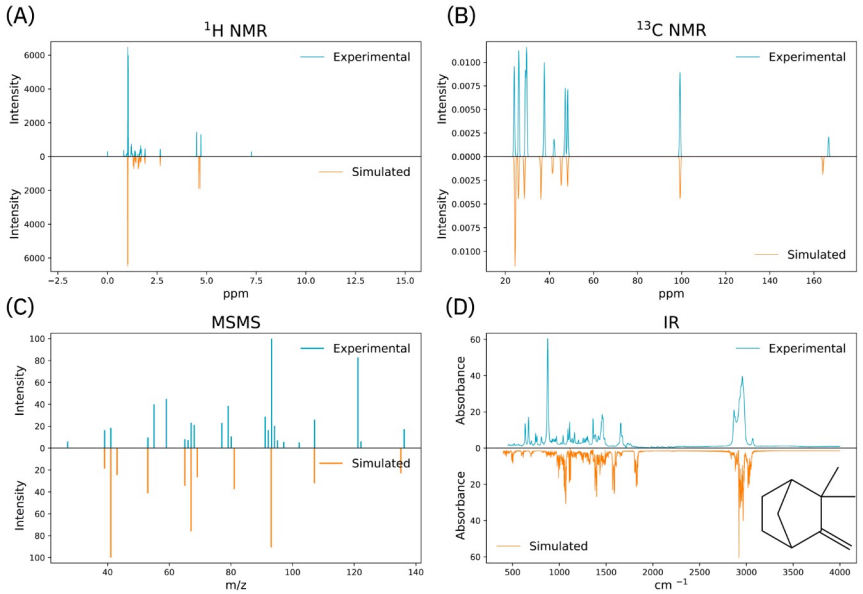


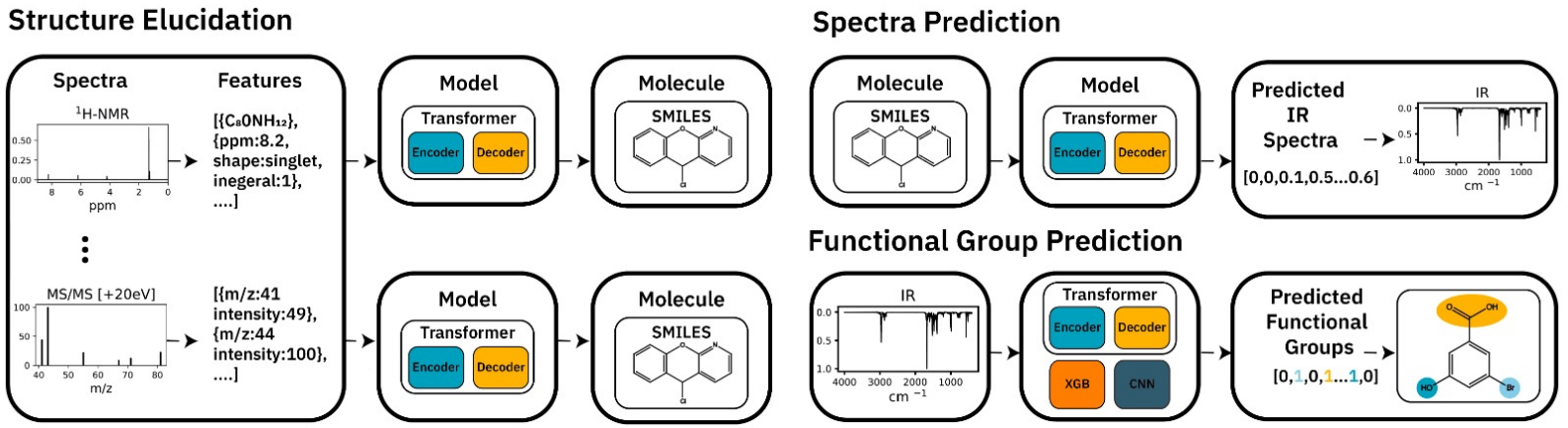
Figure 1: Overall workflow: Molecules are extracted from reaction data (USPTO), filtered to only contain certain atom types as well as minimum and maximum molecule size, then for each molecule the corresponding spectra are simulated resulting in a dataset of spectra for 790k molecules.

Modality	Subtype	Data Description
IR	Spectrum	Vector of size 1.800
^{1}H-NMR	Spectrum	Vector of size 10.000
	Annotated Spectrum	Start, End, Centroid, Integration and Type of each peak
^{13}C-NMR	Spectrum	Vector of size 10.000
	Annotated Spectrum	Centroid and Intensity of each peak
HSQC-NMR	Spectrum	Matrix: 512x512
	Annotated Spectrum	X, Y coordinates and integration of each peak
Positive MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak
Negative MS/MS	Spectrum	m/z & Intensity of each peak
	m/z Annotations	Chemical formula corresponding to the m/z of each peak



Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry, NeurIPS 2024

multi-modal spectra dataset, NeurIPS 2024

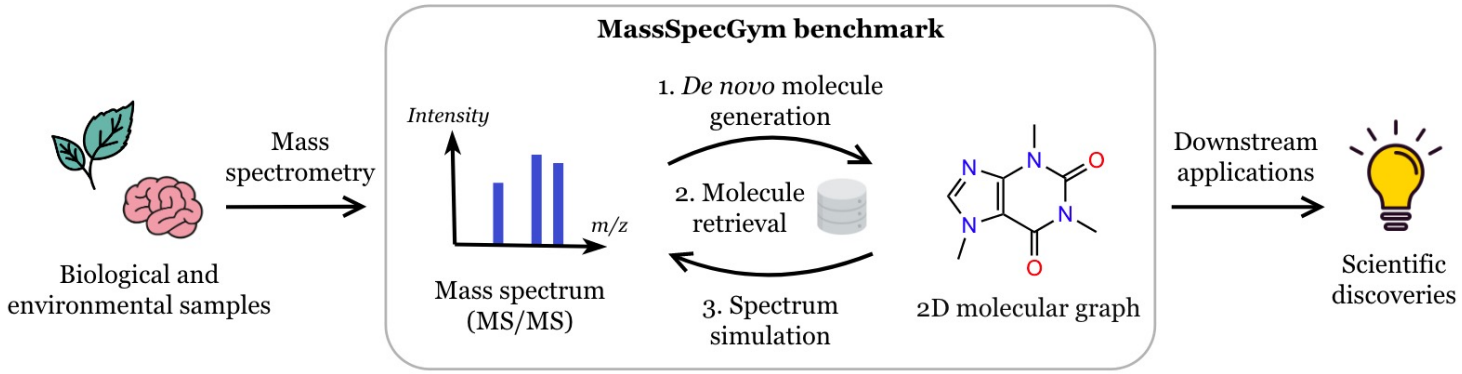


	Top–1%	Top–5%	Top–10%
IR	9.97 ± 0.46	21.23 ± 0.33	24.01 ± 0.42
MS/MS (CFM-ID, Negative)	20.98 ± 0.23	39.32 ± 0.19	44.93 ± 0.29
MS/MS (CFM-ID, Positive)	23.53 ± 0.21	42.59 ± 0.14	47.53 ± 0.31
MS/MS (SCARF, Positive)	1.92 ± 0.11	5.26 ± 0.37	6.81 ± 0.48
MS/MS (ICEBERG, Positive)	15.52 ± 2.10	31.46 ± 3.28	36.22 ± 3.45
^{13}C-NMR	51.95 ± 0.29	70.01 ± 0.21	74.12 ± 0.30
^{1}H-NMR	64.99 ± 0.31	81.94 ± 0.31	84.07 ± 0.32
^{1}H-NMR + ^{13}C-NMR	73.38 ± 0.08	87.94 ± 0.14	89.98 ± 0.16

Spectrum	Cosine Similarity	Token Accuracy
IR	23.91	13.55
MS/MS (Positive) [10 eV]	83.94	31.58
MS/MS (Positive) [20 eV]	77.09	11.05
MS/MS (Positive) [40 eV]	66.35	6.94
MS/MS (Negative) [10 eV]	82.87	33.92
MS/MS (Negative) [20 eV]	75.86	11.82
MS/MS (Negative) [40 eV]	69.50	8.95
^{13}C-NMR	92.69	35.7
^{1}H-NMR	94.86	17.93

Unraveling Molecular Structure: A Multimodal Spectroscopic Dataset for Chemistry, NeurIPS 2024

MassSpecGym, NeurIPS 2024



Dataset	Spectra	High-quality spectra	Molecules	Split
GNPS [36]	**322K**	104K	16K	✗
MoNA [37]	98K	62K	10K	✗
MassBank [38]	62K	58K	4K	✗
MIST CANOPUS [19]	11K	$\leq$ 11K	$\leq$ 9K	✓
MassSpecGym (ours)	231K	**231K**	**29K**	✓

Figure 1: **MassSpecGym provides three challenges for benchmarking the discovery and identification of new molecules from MS/MS spectra**. The provided challenges abstract the process of scientific discovery from biological and environmental samples into well-defined machine learning problems.

Table 2: **Baseline results for the *de novo* molecule generation challenge.** The values in brackets indicate 99.9% confidence intervals upon bootstrapping (20,000 resamples).

	Top-1			Top-10		
	Accuracy ↑	MCES ↓	Tanimoto ↑	Accuracy ↑	MCES ↓	Tanimoto ↑
Random chemical generation	0.00	**28.59 (28.33-28.84)**	0.07 (0.07 - 0.07)	0.00	25.72 (25.49-25.95)	0.10 (0.10 - 0.10)
SMILES Transformer	0.00	53.80 (52.95-54.61)	0.07 (0.07 - 0.08)	0.00	21.97 (21.79-22.16)	**0.17 (0.17 - 0.17)**
SELFIES Transformer	0.00	33.28 (33.00-33.57)	**0.10 (0.10 - 0.10)**	0.00	**21.84 (21.67-22.00)**	0.15 (0.15 - 0.15)
Bonus chemical formulae challenge						
SMILES Transformer	0.00	79.39 (78.64-80.08)	0.03 (0.03 - 0.04)	0.00	52.13 (51.45-52.81)	0.10 (0.09 - 0.10)
SELFIES Transformer	0.00	38.88 (38.57-39.20)	**0.08 (0.08 - 0.08)**	0.00	26.87 (26.66-27.11)	**0.13 (0.13 - 0.13)**
Random chemical generation	0.00	**21.11 (20.97-21.26)**	**0.08 (0.08 - 0.08)**	0.00	**18.25 (18.14-18.35)**	0.11 (0.11 - 0.11)

MassSpecGym: A benchmark for the discovery and identification of molecules, NeurIPS 2024

Molpuzzle, NeurIPS 2024

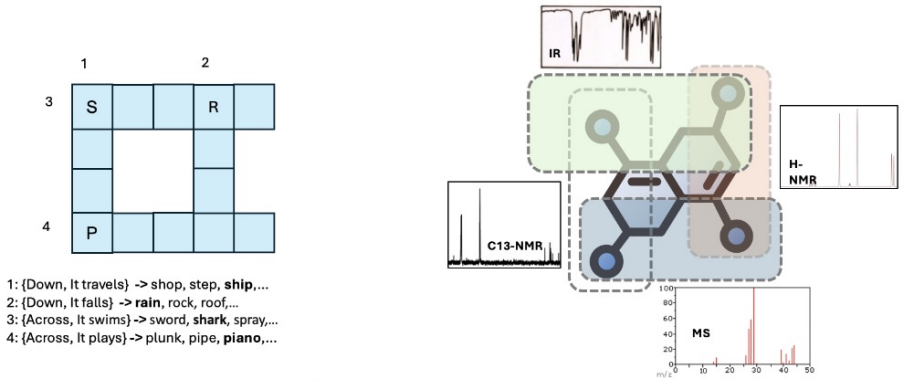


Figure 1: A crossword puzzle (left), and a molecular structure elucidation puzzle (right)

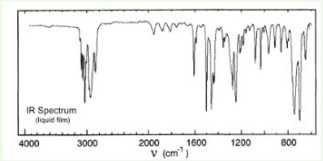
1. Identify molecule substructures based on molecule formula	**2. Refine the substructure pools based on Spectrum images.**	**3. Select fragments from the pools and assemble molecule iteratively**
Prompt: As an expert organic chemist, your task is to analyze the chemical formula C6H10O6 and determine the potential molecular structures and the degree of unsaturation. Utilize your knowledge to systematically explore and identify plausible molecular substructure.	 Prompt: As an expert in organic chemistry, you are tasked with analyzing potential molecular structures derived from **IR spectral data**. Given the molecular formula and an initial set of potential fragment SMILES identified, your objective is to explore and systematically determine plausible molecular substructure that are consistent with the IR spectral data.	**Initial selection:** Prompt: Selected one fragment from the list of SMILES for the Initial structure for molecular construction: Identify one specific fragment from the **[pool of fragments]** provided: ensuring it's consistent with both**[C13-NMR]** and **[H-NMR].** **Iteration:** Prompt: Select one fragment from the provided list of SMILES to add to the current molecule. Identify a specific fragment from the **[pool of fragments]**: , ensuring it is consistent with both the **[C13-NMR**] and **[H-NMR]** spectra. **End:** when run out of heavy atoms.
Answer: Carboxylic Acid (Yes) degree of unsaturation = 2	Answer: ["C(=O)O", "C(=O)OC", "C=O", "CO", "C1CO1"]	Answer: C1C(C(C(C(O1)O)O)O)C(=O)O
(a). The Initial Stage	(b). The Second Stage	(c). The Final Stage

Figure 2: Examples of QA pairs in the 3 stages of MolPuzzle

Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation, NeurIPS 2024

Molpuzzle, NeurIPS 2024

Statistic	Number
Total MolPuzzle Instances	217
Stage-1 QA samples	5,859
- Num. of molecule formula	176
- Max question length	128
- Average question length	94
Stage-2 QA samples	11,501
- Num. of spectrum images	868
- Max question length	340
- Average question length	264
Stage-3 QA samples	6,318
- Maximum Iteration	7
- Max question length	356
- Average question length	238

Figure 3: Statistic of the MolPuzzle dataset

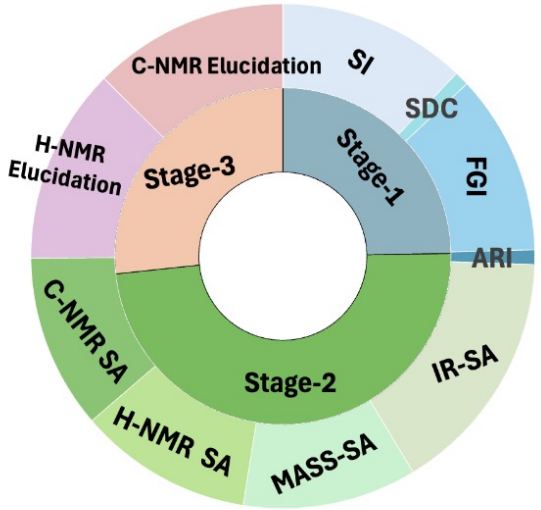


Figure 4: Inner ring: sample distribution in 3 stages Outer ring: sample distribution across categories in each stage. SI: saturation identification, SDC: saturation degree calculation, FGI: functional group identification, ARI: aromatic ring identification, SA: spectrum analysis.

Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation, NeurIPS 2024

Molpuzzle, NeurIPS 2024

Table 1: F1 scores (↑) of individual QA tasks in three stages. The best LLMs results are in bold font. Tasks in stage 1 are SI-Saturation Identification, ARI-Aromatic Ring Identification, FGI-Functional Group Identification, and SDC-Saturation Degree Calculation.

Stage 1 (Molecule Understanding) Tasks				
Method	SI	ARI	FGI	SDC
GPT-4o	**1.00±0.000**	0.943±0.016	0.934±0.005	0.667±0.003
GPT-3.5-turbo	0.451±0.025	0.816±0.017	0.826±0.075	0.5±0.099
Claude-3-opus	0.361±0.009	**0.988±0.015**	**0.934±0.001**	**0.856±0.016**
Galactica-30b	0.826±0.248	0.347±0.000	0.467±0.005	0.000±0.000
Llama3	0.228±0.043	0.696±0.051	0.521±0.003	0.000±0.000
Human	1.00±0.000	1.000±0.000	0.890±0.259	0.851±0.342

Stage 2 (Spectrum Interpretation) Tasks				
Method	IR Interpretation	MASS Interpretation	H-NMR Interpretation	C-NMR Interpretation
GPT-4o	**0.656±0.052**	**0.609±0.042**	**0.618±0.026**	**0.639±0.010**
LLava	0.256±0.026	0.101±0.021	0.118±0.008	0.254±0.015
Human	0.753±0.221	0.730±0.11	0.764±0.169	0.769±0.101

Stage-3 (Molecule Construction) Tasks		
Method	H-NMR Elucidation	C-NMR Elucidation
GPT-4o	**0.524±0.021**	**0.506±0.037**
Llama3	0.341±0.015	0.352±0.017
Human	0.867±0.230	0.730±0.220

Can LLMs Solve Molecule Puzzles? A Multimodal Benchmark for Molecular Structure Elucidation, NeurIPS 2024